Nearly exact tests in factorial experiments using the aligned rank transform

By: Scott J. Richter

Richter, S. J. & Payton, M. E. (1999). Nearly Exact Tests in Factorial Experiments Using the Aligned Rank Transform. *Journal of Applied Statistics*, 26(2), 203-217.

Abstract:

A procedure is studied that uses rank-transformed data to perform exact and estimated exact tests, which is an alternative to the commonly used F-ratio test procedure. First, a common parametric test statistic is computed using rank-transformed data, where two methods of ranking – ranks taken for the original observations and ranks taken after aligning the observations – are studied. Significance is then determined using either the exact permutation distribution of the statistic or an estimate of this distribution based on a random sample of all possible permutations. Simulation studies compare the performance of this method with the normal theory parametric F-test and the traditional rank transform procedure. Power and nominal type I error rates are compared under conditions when normal theory assumptions are satisfied, as well as when these assumptions are violated. The method is studied for a two-factor factorial arrangement of treatments in a completely randomized design and for a split-unit experiment. The power of the tests rivals the parametric F-test when normal theory assumptions are satisfied, and is usually superior when normal theory assumptions are not satisfied. Based on the evidence of this study, the exact aligned rank procedure appears to be the overall best choice for performing tests in a general factorial experiment.

Keywords: exact aligned rank procedure | F-ration test procedure | exact aligned test procedure

Article:

1. Introduction

In experiments to determine if one or more factors have an effect on a response, the researcher typically can choose between one of two classes of analysis: para-metric and non-para metric analysis. Parametric procedures exist for simple and for complex experiments, but the validity of inferences made using these procedures depends on a set of unknown assumptions. The most common of these in the analysis of designed experiments is the assumption of normally distributed populations with equal variances. However, it is generally unknown to what extent

the validity of the inferences suffers when the assumptions are not satisfied. In contrast, m any non-para metric procedures require less stringent assumptions, such as independent samples and observations, which can often be controlled by the experimenter. Furthermore, most of these methods depend on the exact permutation distribution of the test statistic for making inferences. However, as a result of the complexity of deriving the exact sampling distributions when sample sizes are large, most non-para metric methods rely on the asymptotic distribution of the test statistic. In addition, there exist few non-para metric procedures for analyzing complex experimental designs, and most of those that do exist are very limited in application.

Conover and Iman (1976) addressed this situation, by proposing the procedure of performing parametric procedures on the ranks of the data when it was suspected that the parametric assumptions were violated. Many studies of the 'rank transform' procedure, however, have shown it to be non-robust and lacking in power in some situations – most notably, in experiments w here interaction is present (see Akritas, 1990; Blair *et al*., 1987; Sawilowsky *et al*., 1989; Thompson & Ammann, 1990).

An adjustment to the usual rank transform, known as 'ranking after alignment', was first proposed by Hodge s and Lehmann (1962). This adjustment has been found to make the rank transform procedure more robust and more powerful in some situations, especially in designs with interaction. However, asymptotic sampling distributions are still used for tests of significance, and very few studies of the small-sample properties are available. Fawcett and Salter (1984) and Groggel (1987) investigated the aligned rank procedure for testing main effects in a randomized block design. Conover and Iman (1976) examined the aligned rank procedure for testing for interaction in a two-factor factorial experiment, using small effect magnitudes. Higgins *et al*. (1990) and H igg ins and Tashtoush (1994) considered the aligned rank procedure for testing main effects and interaction in a two-factor factorial experiment, and for testing main and subunit effects and interaction in a split-unit experiment.

In this paper, the performances of the usual rank transform and the aligned rank transform are investigated when the exact permutation distribution of the sampling distribution of the test statistic is used. Simulation studies compare the performances of these methods with that of the parametric *F*-ratio test procedure when testing main effects and interaction in factorial and split-unit experiments. Finally, a numerical example is presented to illustrate the implementation of the method.

2. Estimating exact distributions

For complex designs with large sample sizes, the exact distribution of the test statistic will be estimated based on a random sample of all possible permutations of the data. This method was first proposed by D ass (1957) as 'the most logical' way to obtain an approximation to Fisher's method of randomization, and tests based on this method of determining significance have become known as 'randomization tests' (Edgington, 1995; Manly, 1991). This technique, when applied to the actual observations, has the somewhat undesirable property that a possibly unique sampling distribution must be constructed for each set of data. In addition, two researchers performing a randomization test independently on the same set of data would probably obtain slightly different *p*-values. For a large random sample (say 20 000) of permutations, however, it

is unlikely that two independent tests would arrive at different conclusions with regard to significance. For example, for estimating the cumulative probability associated with the 95th percentile of a sampling distribution based on a random sample of 20000 permutations, the expected error of estimation, with 99% confidence, would be about 0.004, or 0.4%. Thus, very precise estimates of the exact critical values of the sampling distribution can be obtained. When applied to rank-transformed data, however, a unique sampling distribution would need to be derived only for each possible sample size. Thus, it is possible to create tables of critical values, given a particular sample size.

3. Simulation study for a completely randomized two-factor factorial experiment

3.1 Procedure

Simulated data sets were generated to examine the performance of the three methods: the parametric *F* -test procedure (FT), the exact rank transform test procedure (RT) and the exact aligned rank transform test procedure (ART). The following model was used to generate the observations:

$$Y_{ijk} = \mu + A_i + B_j + \mathrm{AB}_{ij} + e_{ijk}$$

Here, A_i is the effect of the *i*th level of treatment *A*, *i* 5 1, 2, 3, 4; B_j is the effect of the *j*th level of treatment *B*, *j* 5 1, 2, 3; AB_{ij} is the effect of the interaction between the *i*th level of factor *A* and the *j*th level of factor *B;* and e_{ijk} is the random error effect, *k* 5 1, 2, . . . , *n*. Although most simulations investigated models for samples of size *n* 5 2, some models were also simulated with *n* 5 5 and *n* 5 10.

For the A RT, observations were aligned in the following manner: when testing interaction, an aligned observation was

$$AY_{ijk} = Y_{ijk} - (\text{sample mean})_i - (\text{sample mean})_j$$

when testing for main effects, an aligned observation for testing effect *A* was

$$Ay_{ijk} = Y_{ijk} - (\text{sample mean})_j$$

and, for testing effect *B*, the aligned observation was

$$Ay_{ijk} = Y_{ijk} - (\text{sample mean})_i$$

Standard nor m al and exponential (*l* 5 3) distributions were used to model the error distributions. Effect sizes (denoted by *c* in the tabulated results) are in standard deviation units, and range in magnitude from 0.5 (very small) to 3.5 (very large). In addition, models with variance heterogeneity were investigated. In all cases, the reported variance ratio rep resents the ratio of the largest to the smallest variance. Critical values for both rank tests were estimated by

calculating the value of the test statistic for a random sample of 20 000 permutations of the ranks of the data. 10000 samples were generated and the proportion of test statistic values greater than or equal to the critical values for the respective sampling distributions was calculated. Thus, for estimating a nominal type I error rate of 0.05, the maxim um error of estimation is 0.0056, with 99% confidence (values outside of this range are in **bold** in the tables that follow).

3.2 Results

First, we consider normally distributed errors with equal variances (see Tables 1 and 2). The ART consistently showed power almost equal to that of the FT. The ART often had slightly inflated, nominal type I error rates, but the inflation was never severe and did not appear to be affected by the magnitude of the modeled effects. The RT tended to com pare favorably in most cases, but showed poor power when both main effects and interaction were present in the model—especially for testing interaction. In addition, for all models, the RT h ad nominal type I error rates that inflated as the magnitude of the effects increased. For a more detailed study of the performance of the RT when the parametric assumptions are satisfied, see Blair *et al*. (1987).

TABLE 1. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with equal variance

			c			
n	Test for	Test	0.5	1.5	2.5	3.5
2	Factor A	FT	0.210	0.968	1.00	1.00
		RT	0.199	0.942	1.00	1.00
		ART	0.199	0.959	1.00	1.00
	Factor B	FT	0.329	0.999	1.00	1.00
		RT	0.317	0.996	1.00	1.00
		ART	0.319	0.998	1.00	1.00
	Interaction	FT	0.050	0.050	0.050	0.050
		RT	0.054	0.054	0.054	**0.068**
		ART	0.056	0.056	0.056	0.056
10	Interaction	FT	0.049	0.049	0.049	0.049
		RT	0.051	**0.134**	**0.671**	**0.997**
		ART	0.050	0.050	0.050	0.050

Note: A and B main effects present ($a_2 = b_1 = c$, $a_3 = b_2 = -c$; $c = 0$ for all other effects).

TABLE 2. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with equal variance: $n = 2$

		c			
Test for	Test	0.5	1.5	2.5	3.5
Factor A	FT	0.066	0.213	0.527	0.830
	RT	0.066	0.132	0.193	0.218
	ART	0.065	0.153	0.252	0.290
Factor B	FT	0.139	0.780	0.997	1.00
	RT	0.134	0.652	0.940	0.994
	ART	0.140	0.732	0.989	1.00
Interaction	FT	0.069	0.260	0.655	0.931
	RT	0.066	0.153	0.230	0.264
	ART	0.075	0.251	0.617	0.909

Note: A, B and interaction effects present ($ab_{11} = c$, $b_1 = ab_{41} = -c$; $c = 0$ for all other effects).

TABLE 3. Proportion of rejections at $\alpha = 0.05$, identically exponentially distributed errors

			c			
n	Test for	Test	0.5	1.5	2.5	3.5
2	Factor A	FT	0.066	0.246	0.574	0.828
		RT	0.083	0.314	0.621	0.834
		ART	0.086	0.335	0.665	0.877
	Factor B	FT	0.084	0.386	0.762	0.943
		RT	0.119	0.497	0.825	0.956
		ART	0.113	0.485	0.839	0.966
	Interaction	FT	0.055	0.055	0.055	0.055
		RT	**0.058**	**0.059**	**0.059**	**0.057**
		ART	**0.074**	**0.074**	**0.074**	**0.074**
10	Factor A	FT	0.172	0.898	1.00	1.00
		RT	0.329	0.985	1.00	1.00
		ART	0.332	0.993	1.00	1.00
	Factor B	FT	0.251	0.977	1.00	1.00
		RT	0.477	0.999	1.00	1.00
		ART	0.463	1.00	1.00	1.00
	Interaction	FT	0.048	0.048	0.048	0.048
		RT	0.053	**0.060**	**0.078**	**0.121**
		ART	**0.061**	**0.061**	**0.061**	**0.061**

Note: A and B main effects present ($a_2 = b_1 = c$, $a_3 = b_2 = -c$; $c = 0$ for all other effects).

Next, we consider exponentially distributed error s (see Tables 3 and 4). Both rank tests had superior power relative to the FT. A notable exception was the model which had both main effects and interaction present, where the RT again had less power for testing interaction than in other models. Although the power of the RT was about the same as that of the FT for most models (except when effect magnitudes became very large, where the FT usually had more power), it was still outperformed by the ART. Interestingly, for small sample sizes (*n* 5 2 observations per cell), when the error distributions were non-nor mal, the nominal type I error rates for the RT did not show a tendency to inflate as the magnitudes of the effects increased.

Finally, we consider normally distributed errors with unequal variances (see Tables 5 and 6). This was a much more serious problem than the lack of normality. The power for all methods was less than was found in the equal variance case, and this decrease in power became more severe as the degree of heterogeneity between variances increased. However, both rank tests consistently outperformed the FT in the power category, except for the RT in the previously discussed model. However, the FT did often exhibit slightly higher power for very small effect magnitudes. In addition, the ART usually h ad more power for testing interaction than did the RT. Examination of nominal type I error rates for testing interaction when none was modeled revealed that these rates were inflated for all three methods, with more severe inflation occur ring when the variances differed more. This indicated that variance heterogeneity actually tended to be falsely interpreted as interaction more often than would be expected. The ART seemed to be the most sensitive to this false interaction, which is not surprising, because the alignment procedure isolates the effect of interaction; next most sensitive was the FT and then the RT. Thus, it is not surprising that the A RT showed more power when interaction was actually modeled. The RT was the least sensitive to the presence of interaction.

The problem of nominal type I error rate inflation was not limited only to the test for interaction, however. When only one main effect was modeled along with an interaction effect, the nominal type I error rates for testing the unmodeled main effect were also inflated for all methods. Thus, it is apparent that variance heterogeneity can produce very erratic behavior in the analysis.

TABLE 4. Proportion of rejections at $\alpha = 0.05$, identically exponentially distributed errors

			c			
n	Test for	Test	0.5	1.5	2.5	3.5
2	Factor A	FT	0.049	0.063	0.097	0.154
		RT	0.054	0.073	0.094	0.121
		ART	0.057	0.080	0.113	0.151
	Factor B	FT	0.057	0.155	0.362	0.610
		RT	0.073	0.224	0.405	0.576
		ART	0.072	0.208	0.420	0.634
	Interaction	FT	0.058	0.075	0.113	0.186
		RT	0.059	0.082	0.109	0.142
		ART	0.076	0.100	0.153	0.234
10	Factor A	FT	0.059	0.167	0.412	0.707
		RT	0.077	0.238	0.443	0.616
		ART	0.075	0.268	0.549	0.774
	Factor B	FT	0.113	0.638	0.961	1.00
		RT	0.200	0.832	0.986	1.00
		ART	0.185	0.841	0.992	1.00
	Interaction	FT	0.065	0.227	0.592	0.891
		RT	0.089	0.335	0.634	0.836
		ART	0.091	0.412	0.846	0.984

Note: A, B and interaction effects present ($ab_{11} = c$, $b_1 = ab_{41} = -c$; $c = 0$ for all other effects).

TABLE 5. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with unequal variance: $n = 2$

		c			
Test for	Test	0.5	1.5	2.5	3.5
Factor A	FT	0.108	0.218	0.475	0.753
	RT	0.096	0.280	0.562	0.802
	ART	0.097	0.279	0.613	0.874
Factor B	FT	0.108	0.313	0.651	0.887
	RT	0.102	0.380	0.718	0.914
	ART	0.105	0.406	0.757	0.945
Interaction	FT	**0.113**	**0.113**	**0.113**	**0.113**
	RT	**0.080**	**0.099**	**0.110**	**0.111**
	ART	**0.134**	**0.134**	**0.134**	**0.134**

Notes: Ratio of largest to smallest variance, 30 : 1. A and B main effects present ($a_2 = b_1 = c$, $a_3 = b_2 = -c$; $c = 0$ for all other effects).

TABLE 6. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with unequal variance: $n = 2$

		c			
Test for	Test	0.5	1.5	2.5	3.5
Factor A	FT	0.097	0.110	0.132	0.167
	RT	0.076	0.090	0.115	0.146
	ART	0.082	0.093	0.118	0.144
Factor B	FT	0.090	0.160	0.291	0.481
	RT	0.075	0.144	0.275	0.455
	ART	0.075	0.153	0.302	0.506
Interaction	FT	0.117	0.132	0.164	0.211
	RT	0.078	0.086	0.110	0.140
	ART	0.135	0.157	0.193	0.248

Notes: Ratio of largest to smallest variance, 30 : 1. A, B and interaction effects present ($ab_{11} = c$, $b_1 = ab_{41} = -c$; $c = 0$ for all other effects).

4. Simulation study for a split-unit experiment

4.1 Procedure

Simulated data sets were generated to examine the performance of the three methods. A split-unit experiment with main units in a randomized complete block design was considered. The following model was used to generate the observations:

$$Y_{ijk} = B_i + M_j + \mathrm{BM}_{ij} + S_k + \mathrm{SM}_{jk} + E_{ijk}$$

Here, B_i is the random effect of the ith block, $i = 1, 2, 3$; M_j is the fixed effect of the jth level of the main unit treatment, $j = 1, 2, 3, 4$; BM_{ij} is the random effect of the interaction between the ith block and the jth level of the main unit treatment; S_k is the fixed effect of the kth level of the sub unit treatment, $k = 1, 2, 3$; SM_{jk} is the fixed effect of the interaction between the jth level of the subunit treatment with the kth level of the main unit treatment; and E_{ijk} is the random subunit error effect.

The random effect BM_{ij} was used as the error to test for the effect of the main unit treatment, while the random effect E_{ijk} was used as the error to test both the subunit treatment effect S_k and the interaction effect SM_{jk}. Standard nor mal (both with homogeneous and heterogeneous variances), exponential ($\mu = 3$) and uniform [- 3, 3] distributions were used to model the error distributions. 10000 samples were generated, and the proportion of test statistic values greater than or equal to the critical values for the respective sampling distributions was calculated.

For the aligned rank procedure, three different methods of aligning were used, depending on the effect being tested. For testing the main unit treatment effect, the observations were aligned by subtracting estimates of both block and subunit treatment effects. For testing the subunit

treatment effect, estimates of both block and main unit treatment effects were subtracted from each observation. Finally, for testing the interaction, the observations were aligned by subtracting block, main unit and subunit effect estimates.

Once again, in each case where unequal error variances were modeled, the reported ratio represents the ratio of the largest to the smallest variance.

4.2 Results

First, we consider normally distributed main unit and subunit errors (see Tables 7 and 8). In this situation, all random effects were modeled as identically distributed, standard normal distributions. The performance observed for each of the three methods was almost identical to that found in the previous study of the two-way layout in a completely randomized design. Both rank tests consistently exhibited power almost equal to that of the FT. As in the completely randomized case, the RT again showed poor power for testing interaction when both main and subunit main effects and interaction were present in the model. When only main and subunit effects were in the model, the RT again exhibited type I error rates that inflated as the magnitude of the effects increased. However, this behavior was not as evident for other models.

Next, we consider exponentially distributed errors (see Tables 9 and 10). When the sub unit error effect was exponentially distributed, both the rank tests had more power than did the FT for all models. When all the fixed effects were in the model, the power of the A RT was clearly superior to those of the other two tests, although the drop-o þ in power for the RT was not as severe as had been observed in previous situations.

TABLE 7. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with equal variance

		c			
Test for	Test	0.5	1.5	2.5	3.5
MU Trt	FT	0.088	0.474	0.900	0.994
	RT	0.091	0.467	0.889	0.993
	ART	0.096	0.481	0.897	0.993
SU Trt	FT	0.500	1.00	1.00	1.00
	RT	0.449	1.00	1.00	1.00
	ART	0.473	1.00	1.00	1.00
Interaction	FT	0.049	0.049	0.049	0.049
	RT	0.046	0.047	**0.077**	**0.148**
	ART	0.049	0.049	0.049	0.049

Notes: MU (main unit) and SU (sub unit), main effects present ($m_2 = s_1 = c$, $m_3 = s_2 = -c$; $c = 0$ for all other effects).

TABLE 8. Proportion of rejections at $\alpha = 0.05$, normally distributed errors with equal variance

		c			
Test for	Test	0.5	1.5	2.5	3.5
MU Trt	FT	0.052	0.087	0.168	0.298
	RT	0.057	0.078	0.114	0.146
	ART	0.058	0.087	0.123	0.155
SU Trt	FT	0.187	0.942	1.00	1.00
	RT	0.168	0.875	0.998	1.00
	ART	0.179	0.911	1.00	1.00
Interaction	FT	0.079	0.416	0.894	0.997
	RT	0.070	0.269	0.497	0.642
	ART	0.075	0.383	0.850	0.991

Note: MU, SU main effects and interaction effect present ($ms_{11} = -c$, $s_1 = ms_{41} = c$; $c = 0$ for all other effects).

TABLE 9. Proportion of rejections at $\alpha = 0.05$, exponentially distributed subunit errors, normally distributed block effect and main unit errors

		c			
Test for	Test	0.5	1.5	2.5	3.5
MU Trt	FT	0.066	0.198	0.470	0.748
	RT	0.074	0.234	0.513	0.770
	ART	0.074	0.240	0.542	0.801
SU Trt	FT	0.095	0.543	0.909	0.989
	RT	0.126	0.657	0.948	0.996
	ART	0.125	0.655	0.952	0.997
Interaction	FT	0.044	0.044	0.044	0.044
	RT	0.049	0.049	0.049	0.055
	ART	0.058	0.058	0.058	0.058

Note: MU and SU main effects present ($m_2 = s_1 = c$, $m_3 = s_2 = -c$; $c = 0$ for all other effects).

Finally, we consider heterogeneous error s (see Tables 11 ± 14). Two cases were considered. One of the errors was modeled as being normally distributed with heterogeneous variances, while the other error was modeled as being normally distributed with homogeneous variances. In each case, the block effect was modeled as having a standard normal distribution. For all models, the ratio between the largest and the smallest variances was considered to be 30 : 1 (very large). As in the completely randomized case, unequal error variances turned out to be a more serious problem than was the lack of normality. However, while the performance of the rank tests was generally better than that of the FT in the completely randomized case, the results were mixed in the split-unit case.

TABLE 10. Proportion of rejections at $\alpha = 0.05$, exponentially distributed subunit errors, normally distributed block effect and main unit errors

		c			
Test for	Test	0.5	1.5	2.5	3.0
MU Trt	FT	0.054	0.068	0.096	0.138
	RT	0.055	0.070	0.094	0.120
	ART	0.056	0.074	0.098	0.132
SU Trt	FT	0.061	0.220	0.518	0.778
	RT	0.076	0.282	0.574	0.778
	ART	0.076	0.274	0.582	0.805
Interaction	FT	0.050	0.080	0.160	0.288
	RT	0.055	0.094	0.155	0.227
	ART	0.064	0.105	0.198	0.345

Note: MU, SU main effects and interaction effect present ($ms_{11} = -c$, $s_1 = ms_{41} = c$; $c = 0$ for all other effects).

TABLE 11. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, unequal main unit error variances

		c				
Test for	Test	0.0	0.5	1.5	2.5	3.5
MU Trt	FT	**0.083**	0.088	0.130	0.223	0.366
	RT	**0.090**	0.095	0.151	0.257	0.405
	ART	**0.084**	0.090	0.142	0.258	0.407
SU Trt	FT	0.050	0.509	1.00	1.00	1.00
	RT	0.056	0.422	1.00	1.00	1.00
	ART	0.050	0.440	1.00	1.00	1.00
Interaction	FT	0.052	0.052	0.052	0.052	0.052
	RT	0.051	0.057	**0.080**	**0.107**	**0.120**
	ART	0.050	0.050	0.050	0.050	0.050

Notes: Ratio of largest to smallest variance, 30:1. MU and SU main effects present ($m_2 = s_1 = c$, $m_3 = s_2 = -c$; $c = 0$ for all other effects).

The power of all tests was lower when the main units had heterogeneous variances, and the power reduced as the degree of heterogeneity increased. When only main unit and subunit treatment effects were present, the rank tests exhibited better power for testing for main unit treatment effects, but slightly less power for testing for subunit treatment effects. In addition, the RT had nominal type I error rates that increased steadily with increasing effect magnitudes. When all the effects were present, the FT exhibited the best power, with the ART close behind and the RT a distant third.

The rank tests performed consistently better than did the FT when the subunit error effect had unequal variances. When the ratio of the largest to the smallest variance was 30 : 1, the rank tests

exhibited more power. For all the methods, there was also a slight nominal type I error rate inflation for testing the interaction effect, which became more severe as the variance ratio increased. Surprisingly, the RT showed less inflation than did either the FT or the ART. When only both main and subunit effects were modeled, the rank tests were much more powerful, with some nominal type I error rate inflation for testing interaction evident for all the methods. However, while the FT and the ART nominal rates remained constant as the magnitude of the effects increased, the RT showed its familiar inflation as an increasing function of effect magnitude. When all the fixed effects were in the model, the ART exhibited much more power than did the other two methods for testing interaction.

TABLE 12. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, unequal main unit error variances

		c			
Test for	Test	0.5	1.5	2.5	3.5
MU Trt	FT	0.084	0.088	0.097	0.109
	RT	0.091	0.092	0.094	0.101
	ART	0.085	0.087	0.094	0.103
SU Trt	FT	0.194	0.936	1.00	1.00
	RT	0.133	0.691	0.969	1.00
	ART	0.144	0.777	0.991	1.00
Interaction	FT	0.082	0.421	0.890	0.996
	RT	0.067	0.152	0.302	0.458
	ART	0.070	0.307	0.735	0.947

Notes: Ratio of largest to smallest variance, 30 : 1. MU, SU main effects and interaction effect present ($ms_{11} = -c, s_1 = ms_{41} = c, c = 0$ for all other effects).

Investigation of the nominal type I error rates when the main or subunit variances were unequal revealed a problem of inflated nominal type I error rates similar to that of the completely randomized experiment (see Tables 13 and 14). When the main unit variances were heterogeneous, the nominal type I error rates for testing the main unit treatment effects were often larger than expected. When the subunit variances were heterogeneous, the nominal type I error rates for testing for subunit treatment and interaction effects were always inflated. However, heterogeneous main unit variances did not adversely affect the nominal levels of the subunit tests, and vice versa. Once again, the inflation of the nominal rates for the RT was often a function of the magnitude of the modeled effects, while the inflation of the nominal rates for the FT and the ART seemed to be independent of the effect magnitude. This again indicates that, when error variances are heterogeneous, test results may be misleading, especially when testing for interaction. This was not a problem when one of the underlying populations was skewed (exponentially distributed).

TABLE 13. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, unequal subunit error variances

		c				
Test for	Test	0.0	0.5	1.5	2.5	3.5
MU Trt	FT	0.052	0.063	0.155	0.350	0.619
	RT	0.055	0.070	0.184	0.389	0.625
	ART	0.052	0.067	0.191	0.437	0.701
SU Trt	FT	0.074	0.095	0.411	0.911	0.999
	RT	0.073	0.131	0.666	0.985	1.00
	ART	0.068	0.114	0.636	0.984	1.00
Interaction	FT	0.083	0.083	0.083	0.083	0.083
	RT	0.065	0.065	0.074	0.081	0.083
	ART	0.105	0.105	0.105	0.105	0.105

Notes: Ratio of largest to smallest variance, 30:1. MU and SU main effects present ($m_2 = s_1 = c$, $m_3 = s_2 = -c$; $c = 0$ for all other effects).

TABLE 14. Proportion of rejections at $\alpha = 0.05$, normally distributed errors, unequal subunit error variances

		c			
Test for	Test	0.5	1.5	2.5	3.5
MU Trt	FT	0.053	0.059	0.079	0.111
	RT	0.057	0.075	0.095	0.122
	ART	0.054	0.073	0.101	0.135
SU Trt	FT	0.081	0.159	0.370	0.682
	RT	0.090	0.240	0.537	0.816
	ART	0.078	0.210	0.510	0.814
Interaction	FT	0.085	0.102	0.143	0.219
	RT	0.070	0.107	0.170	0.242
	ART	0.108	0.135	0.193	0.294

Notes: Ratio of largest to smallest variance, 30:1. MU, SU main effects and interaction effect present ($ms_{11} = -c$, $s_1 = ms_{41} = c$; $c = 0$ for all other effects).

5. Conclusion and summary

The exact aligned rank procedure appears to be the overall best choice for performing tests in a general factorial experiment. When the error distribution was symmetric and the error variances were homogeneous, the ART was nearly as powerful as was the FT, with an almost negligible difference in power between the two methods. For a skewed error distribution, the ART was clearly more powerful than was the FT. When the error variances were heterogeneous, both methods led to problems with maintaining nominal type I error levels for testing interaction, but the ART showed superior power for detecting main effects and interaction.

Although the results were not as consistent as for the completely randomized case, the exact aligned rank procedure appears to be a viable alternative to the normal theory FT for performing tests in a split-unit factorial design; it is certainly a better choice than is the rank transform method. Once more, when the error distributions were normal and the error variances were homogeneous (situations in which the FT is known to work well), the ART was always nearly as powerful, usually with an almost negligible difference in power between the two methods. For exponential error distributions, the ART was clearly more powerful than the FT. Uniformly distributed errors were also examined for several models. The results were nearly identical to those in the case for normally distributed errors, with the FT having the most power, followed closely by the ART and then the RT. Again, the ART often had slightly inflated nominal type I error rates for testing interaction. When the error variances were heterogeneous, both methods tended to lead to problems with maintaining nominal type I error levels for interaction – although these problems were less severe in the split-unit case – while the ART usually exhibited superior power for detecting main effects.

Although the FT outperformed the ART in some cases, even when parametric assumptions were violated, the ART had superior power in most cases, and tended to enjoy a greater power advantage when it was the more powerful test, especially when the assumptions of normality and homogeneity of variance were violated. Although the simulation results indicate that a non-existent interaction effect can be introduced when error variances are unequal, this phenomenon occurs for the FT and for the ART. Because the analysis is typically performed without the benefit of definite knowledge of the nature of the error variances, and because the ART generally has more power than does the FT when the variances are unequal, the ART seems a logical choice over the FT.

One issue that deserves comment is the choice of estimator used for aligning observations. The mean was used in this study, but an argument could be made for using a more robust measure, especially when the error distribution is skewed. Higgins and Tashtoush (1994) examined the use of the trim med mean and the median, but concluded that the gain in power did not necessarily outweigh the greater ease of implementation of the procedure using the mean. Also, regardless of which estimator of location is used, the performance of the test maybe affected by the properties of that estimator for the underlying error distribution. This may explain the inflated type I error rates observed for samples from skewed distributions, for example, where the mean is probably not the most robust measure of location.

Another issue is the problem of heterogeneous errors, which is generally considered to be a more serious problem than is departure from normality. Other transformations can sometimes be used to lessen the effect of variance heterogeneity but, because the purpose of this study was to improve the performance of the rank transform procedure, additional transformations were not investigated. However, it is possible that an additional transformation could help to alleviate the problem of inflated nominal type I error rates.

6. Example

The following example (Ott, 1993, p. 884) illustrates an experiment conducted to determine the effects of four different pesticides ($A1$, $A2$, $A3$, $A4$) on the yield of fruit from three different varieties ($B1$, $B2$, $B3$) of a citrus tree. Eight trees from each variety were randomly selected from an orchard. The four pesticides were then randomly assigned to two trees of a particular variety and applications were made according to the recommended levels. The yields of fruit, in bushels per tree, were obtained after the test period. These data appear in Table 15.

TABLE 15. Yields, in bushels, of 24 citrus trees, with two trees of each variety randomly assigned to one of four pesticides

	Pesticide (A)				
Variable (B)	1	2	3	4	Mean
1	49	50	43	53	46.88
	39	55	38	48	
2	55	67	53	85	59.25
	41	58	42	73	
3	66	85	69	85	78.25
	68	92	62	99	
Mean	53.00	67.83	51.17	73.83	61.46

TABLE 16. The aligned observations obtained by subtracting the respective variety and pesticide means from the observations in Table 15

	Pesticide (A)			
Variety (B)	1	2	3	4
	− 50.88	− 64.71	− 55.05	− 67.71
1	− 60.88	− 59.71	− 60.05	− 72.71
	− 57.25	− 60.08	− 57.42	− 48.08
2	− 71.25	− 69.08	− 68.42	− 60.08
	− 65.25	− 61.08	− 60.42	− 67.08
3	− 63.25	− 54.08	− 67.42	− 53.08

TABLE 17. Ranks of the aligned observations in Table 16

	Pesticide (A)			
Variety (B)	1	2	3	4
	23	9	20	5
1	12	17	16	1
	19	14.5	18	24
2	2	3	4	14.5
	8	11	13	7
3	10	21	6	22

We will illustrate the test for the interaction effect (*AB*). Subtracting the corresponding row and column means from each observation, we obtain the aligned observations in Table 16.

Ranking these observations, without regard to factor level, we obtain the results shown in Table 17.

Computing the ordinary *F*-ratio statistic, $F = \mathrm{MS}(AB)/\mathrm{M\,S}(E)$, we obtain $F = 82.542/54.271 = 1.52$. Because the 0.9 quantile of this statistic is 2.356, there is insufficient evidence of an interaction effect. It was found that the estimated exact tail quantiles of the aligned rank *F* -ratio statistics were very close to the theoretical *F* distribution (for this example, $F(0.9, 6, 12)$ 5 2.33). Thus, in practice, there will be little difference in using tables of the *F* distribution to determine significance instead of the exact quantiles.

REFERENCES

A KRITAS, M. G. (1990) The rank transform method in some two-factor designs, *Journal of the American Statistical Association*, 85, pp. 73 ± 78.

BLAIR, R. C., SAWILOWSKY, S. S. & H IGGINS, J. J. (1987) Limitations of the rank transform statistic in tests of interaction, *Communications in Statistics: Computation and Simulation*, B16, pp. 1133 - 1145.

C ONOVER, W. J. & IMAN, R. L. (1976) On some alternative procedures using ranks for the analysis of experimental designs, *Communications in Statistics: Theory and Methods*, A5, pp. 1349 - 1368.

D WASS, M. (1957) Modified randomization tests for nonparametric hypotheses, *Annals of Mathematical Statistics*, 28, pp. 181 - 187.

E DGINGTON, E. S. (1995) *Randomization Tests*, 3rd Edn (New York, Marcel Dekker).

FAWCETT, R. F. & SALTER, K. C. (1984) A Monte Carlo study of the *F*-test and three tests based on ranks of treatment effects in randomized block designs, *Communications in Statistics: Simulation and Computation*, B13, pp. 21 3 - 225 .

G ROGGEL, D. J. (1987) A Monte Carlo study of rank tests for block designs, *Communications in Statistics: Simulation and Computation*, 16, pp. 601 ± 620.

H IGGINS, J. J., BLAIR, R. C. & TASHTOUSH, S. (1990) The aligned rank transform procedure, *Proceedings of the 1990 Kansas State University Conference on Applied Statistics in Agriculture*, pp. 18 5 - 195.

H IGGINS, J. J. & TASH TOUSH, S. (1994) An aligned rank transform test for interaction, *Nonlinear World*, 1, pp. 201 - 211.

HODGES, J. L. & L EHMANN, E. L. (1962) Rank methods for combination of independent experiments in analysis of variance, *Annals of Mathematica l Statistics*, 27, pp. 324 - 335.

M ANLY, B. F. J. (1991) *Randomization and Monte Carlo Methods in Biology* (New York, Chapman & Hall).

OTT, R. L. (1993) *An Introduction to Statistical Methods and Data Analysis* (Belmont, C A, Wadsworth).

SAWILOWSKY, S. S., BLAIR, R. C. & H IGGINS, J. J. (1989) An investigation of the type I error and power properties of the rank transform procedure in factorial AN OVA, *Journal of Educational Statistics*, 14, pp. 25 5 - 267 .

THOMPSON, G. L. & AMMANN, L. P. (1990) Efficiencies of interblock rank statistics for repeated measures designs, *Journal of the American Statistical Association*, 85, pp. 519 - 528.